

Exploring Art, Theater and **Sentence Entropy** with Machine Learning





Hello!

I'm **Miles Thompson**

Mostly I'm a data scientist and developer but sometimes I like to pretend I'm an artist ;-)

You can find me at miles@godbright.nz





The great outdoors





Tom Lorenzo

4 February · Hong Kong · 🌐



Anybody wanting to work with me in a project involving Machine Learning / Tensorflow and art?

I am in the urgent need of hiring a Research Assistant / Senior Research Assistant. 13 months contract, with possible extension.

Programming experience is required. The project is in Hong Kong (visa could be sponsored).

Please share with whoever may be interested. Contact me at tomas@laurenzo.net. Thanks!



Some Motivation



Table of Contents

- Context
- Data collection
- Entropy
- Other experiments



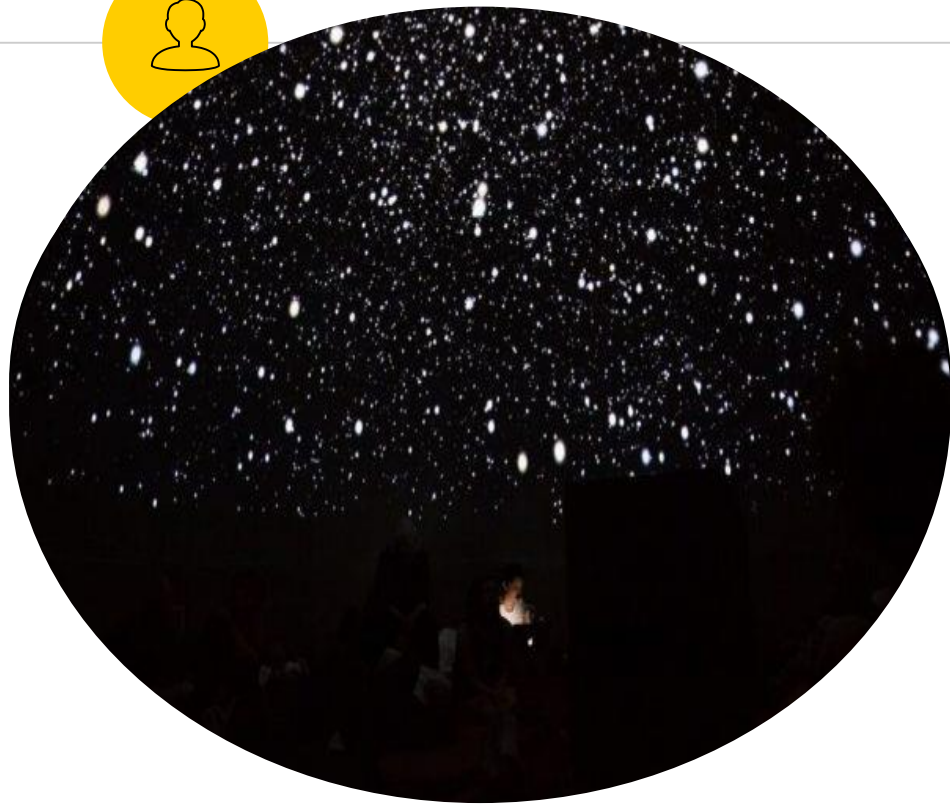
Art

Computers

Algorithms

**Machine
Learning**





About the director, **Annie Dorsen**

Broadway

– “Passing Strange”

Algorithmic Theater

- “Hello Hi There”

- “Yesterday Tomorrow”

- “The Great Outdoors”

“The audience will gather around the flickering projector, with all kinds of text and stories summoned down to us from the “cloud.” .. We gather together to hear the stories not so much from our ancestors as from each other, right now, two seconds ago.”



“



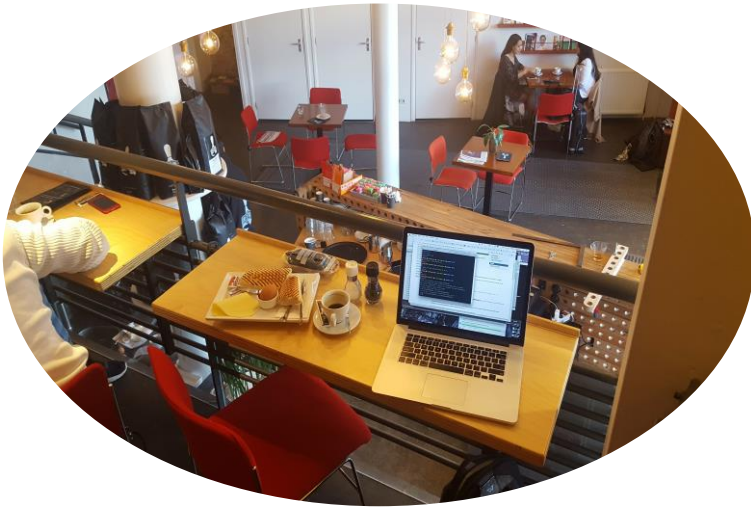
Development process





Some of the **people involved**

- Director
- Music programming
- Video programming
- Text programming
- Dramaturgy
- Our amazing actor/s
- The awesome intern



The theater **process**

- Theater - turns out “it’s a whole thing” ;-)
- Need very rapid iterations (< 1 hour)
- **Solution:** Split code between preparation and script.

Adjust “script” code during rapid iteration, do numeric processing offline.



Data Collection.



Pushshift.io



Directory Contents

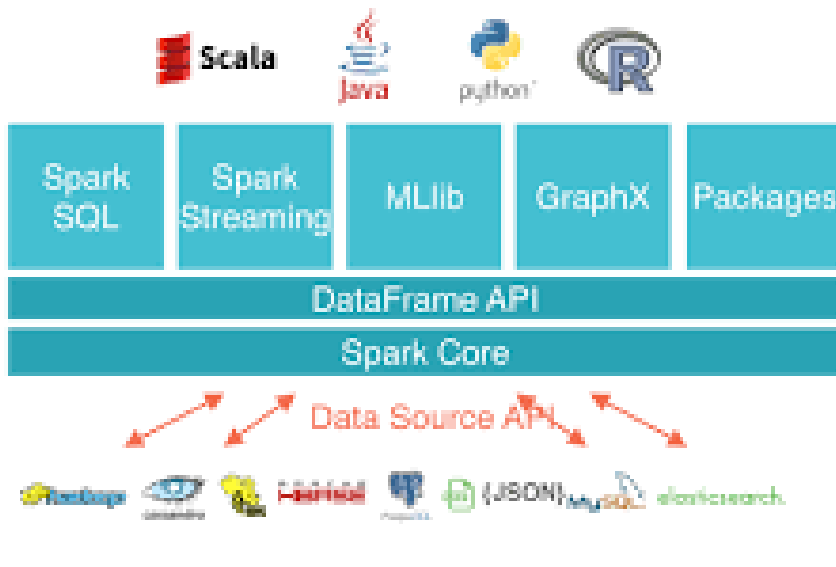
Please consider making a donation (<https://pushshift.io/donations>) if you download a lot of data. This helps offset the costs of my time collecting data and providing bandwidth to make these files available to the public. Thank you!

If you have any questions about the data formats of the files or any other questions, please feel free to contact me at jason@pushshift.io

Filename	Type	Size (bytes)	Date Modified
RC_2017-09.bz2	BZIP2 Compressed Reddit Comments (JSON objects)	7,763,553,912	Oct 6 2017 12:27 PM
RC_2017-10.bz2	BZIP2 Compressed Reddit Comments (JSON objects)	9,963,378,442	Nov 12 2017 5:45 PM
RC_2017-11.bz2	BZIP2 Compressed Reddit Comments (JSON objects)	9,971,136,023	Dec 14 2017 4:01 AM
RC_2017-12.xz	LZMA2 Compressed Reddit Comments (JSON objects)	7,635,154,760	Jan 7 2018 9:17 PM
RC_2018-01.xz	LZMA2 Compressed Reddit Comments (JSON objects)	8,228,348,756	Feb 12 2018 9:02 PM
README.txt	Text File	89	Jul 16 2017 9:38 PM
daily	<Directory> File	<Directory>	Mar 12 2018 9:56 PM
monthComment	Text File	8,500	Feb 10 2018 9:17 PM

Spark / Databricks

Power but \$\$



Databricks

Clusters / job-10872-run-1

job-10872-run-1

Configuration Notebooks (0) Libraries (0) Spark UI Driver Logs Spark Cluster UI - Master

Hostname: HISTORY_SERVER Spark Version: custom:spark-image-c99012b3d17deb5c6ef1355415ebd4e110699435b103a262bc0d7e5a6af7e

Job	Stages	Storage	Environment	Executors	*details	00/0/44			KB
15	6122353838083489490	collect at <console>-40	case class Data(id: Long) val df = sc.parallel...	2016/08/03 0.3 s	4/4				61.5 KB
14	6122353838083489490	collect at <console>-40	case class Data(id: Long) val df = sc.parallel...	2016/08/03 0.4 s	4/4				
13	6122353838083489490	collect at <console>-39	case class Data(id: Long) val df = sc.parallel...	2016/08/03 0.8 s	205/200				61.5 KB
12	6122353838083489490	collect at <console>-39	case class Data(id: Long) val df = sc.parallel...	2016/08/03 0.3 s	4/4				61.5 KB
11	6122353838083489490	collect at <console>-39	case class Data(id: Long) val df = sc.parallel...	2016/08/03 0.4 s	4/4				
10	6122353838083489490	saveAsTable at <console>-36	import java.lang.management.ManagementFactory v...	2016/08/03 1 s	4/4				
9	6122353838083489490	saveAsTable at <console>-37	import java.lang.management.ManagementFactory v...	2016/08/03 0.9 s	4/4				
8	1886757717289855227	show tables	take at OutputAggregator scala:80	2016/08/03 88 ms	1/1				
7	4715541688260152232	first at <console>-35		2016/08/03 36 ms	3/3				
6	4715541688260152232	first at <console>-35		2016/08/03 43 ms	1/1				

Attached: Small Cluster View: Code Permissions Run All Clear Results Schedule Comments Revision history

Transaction counts by card type

```
display(df.groupBy('credit_card_type').count().collect())
```

(8) Spark Jobs

Command took 8.55 seconds --- by job@pctransacts.com at 8/28/2017, 5:06:04 PM on Small Cluster

Std 4

```
1 # Average payment by card type
2 display(df.groupBy('credit_card_type').agg(("total_amount":"avg").alias('avg_payment').collect()))
```

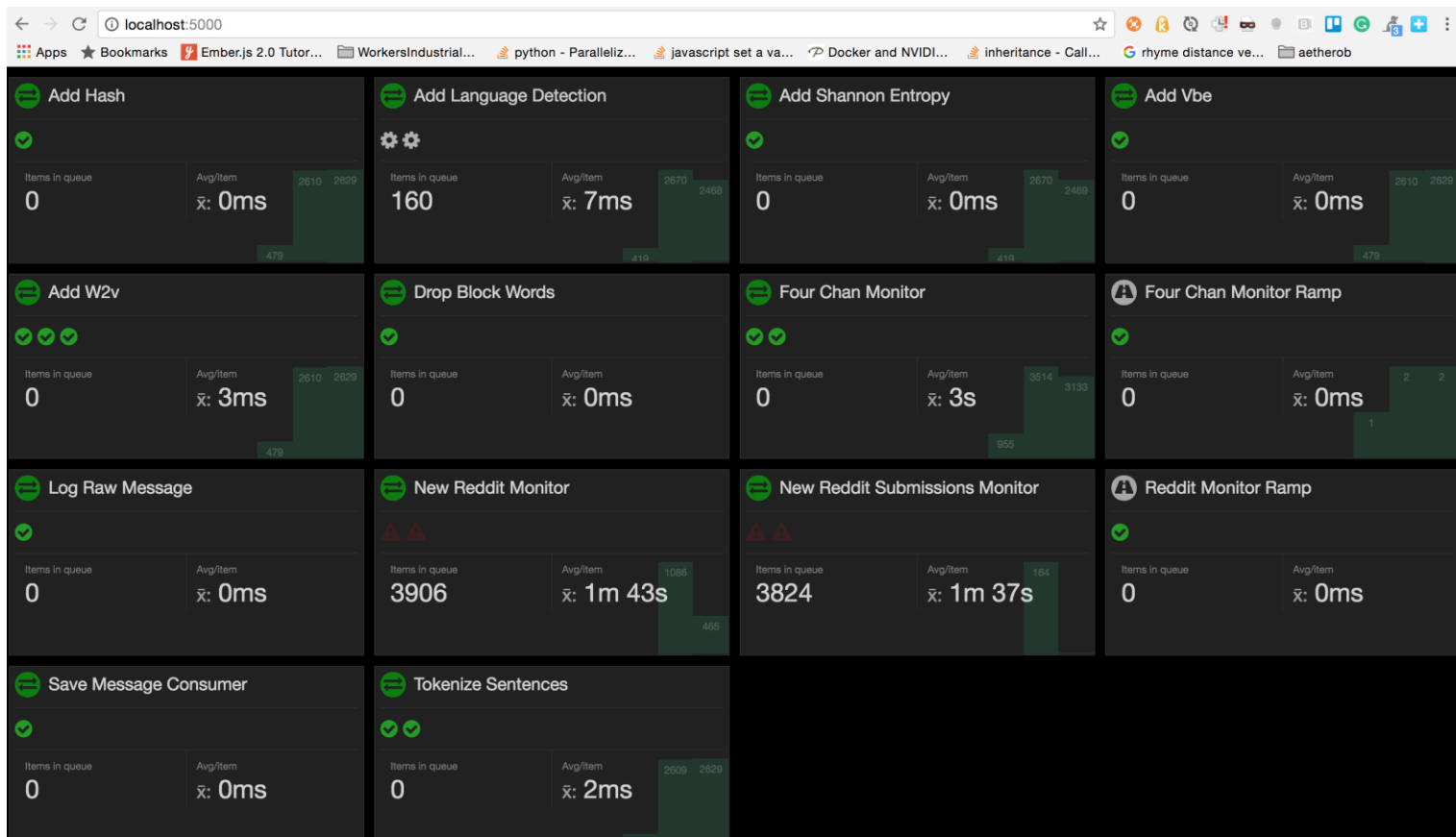
(8) Spark Jobs

Command took 2.03 seconds --- by job@pctransacts.com at 8/28/2017, 5:06:04 PM on Small Cluster

Email Support

Motorway

more home-build but free



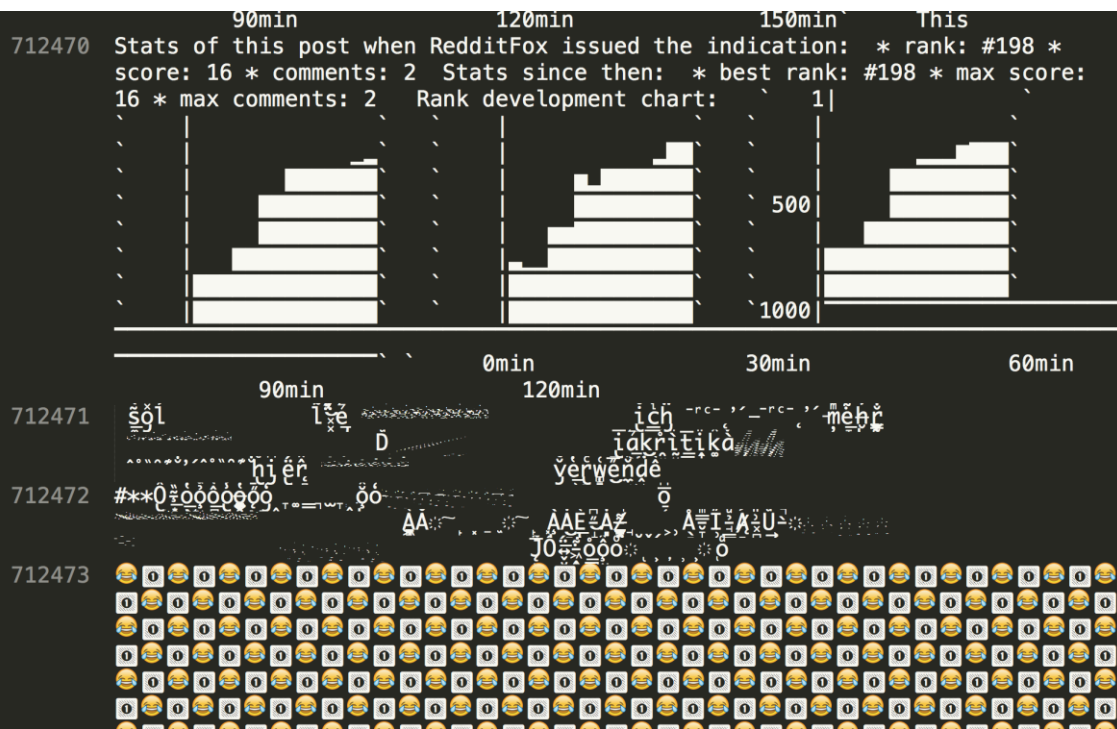
Motorway

more home-build but free

DEMO

A quick dive into some of the text

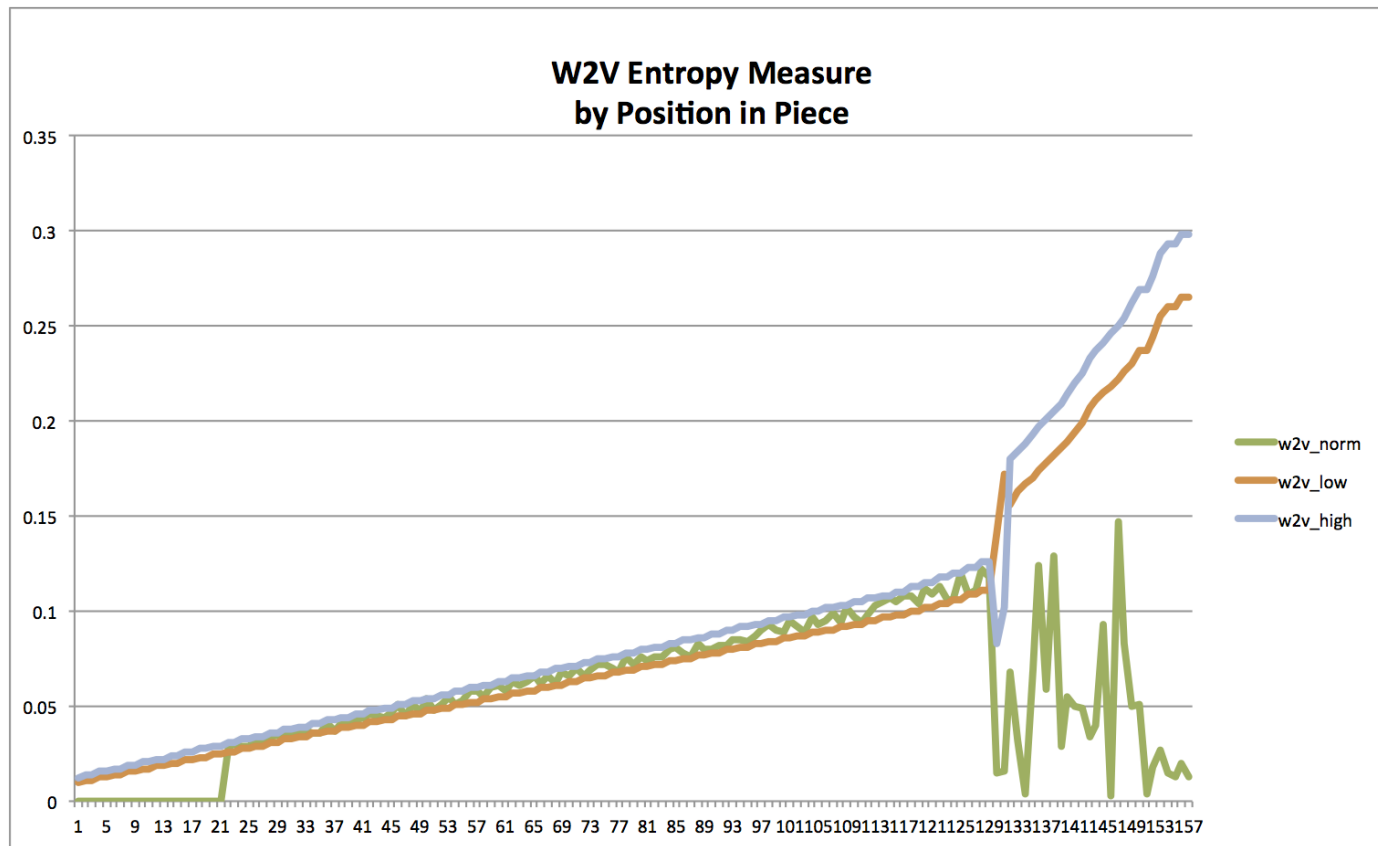
74604 I'm feelin you too
74605 I'm fly, I'm pilot.
74606 I'm from Australia.
74607 I'm from Wisconsin
74608 I'm fucking livid.
74609 I'm happy to help.
74610 I'm imagining this.
74611 I'm in both subs =)
74612 I'm just a packrat!
74613 I'm not ____, but...
74614 I'm not dating her.
74615 I'm not doing this.
74616 I'm not seeing it.
74617 I'm not telling you
74618 I'm okay with this.
74619 I'm part of history
74620 I'm pot committed.
74621 I'm pretty sure IEM



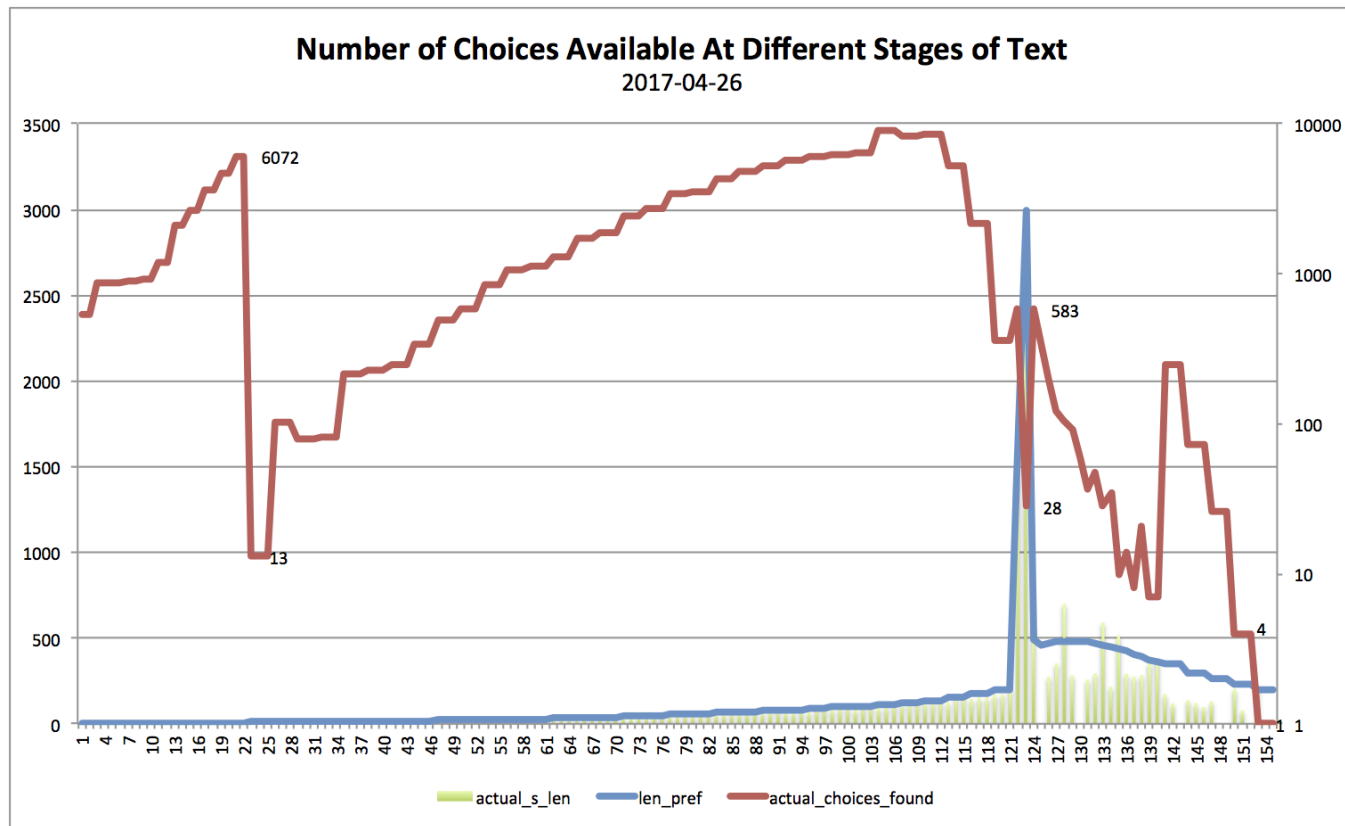
Entropy



The Arc of the piece



The Arc of the piece



Lowest entropy

- 1 Yes
- 2 Yep.
- 3 No.
- 4 Same
- 5 Yeah
- 6 Why?
- 7 Yeah.
- 8 How?
- 9 Yes!
- 10 Wow.
- 11 Agreed.
- 12 Good bot
- 13 Right?
- 14 Holy shit.
- 15 Thank you!

Lowish entropy

```
68 That's not how any of this works.  
69 Thanks guys that's good news. Back to it!  
70 This is REALLY awesome. I love your line work!  
71 Why are you paying the least then? :P  
72 They'll find someone to replace ya. That's how the world works  
73 Is only joke. You don't hef to get mad.  
74 Seeing the signal would have been useful.  
75 I love cats! They always cheer me up :)
```

Middle entropy / More personal, long stories

to do.

123 I recently told my boyfriend that I was diagnosed with severe depression. The thing is, he's trying to compare notes with how he dealt with his and mine. I'm not one to invalidate what he felt over that but I feel like he's not getting that it's an illness. I've had it for a long time already and only recently had myself checked. It was manageable before but it only took its toll on me because I currently don't have a job and the companies I've applied to both local and international rejected my application. He said that I have a lot of unresolved issues to even let my depression get this far in my life. I've had it since I was in college and I'm in my late twenties already. Thing is, I'm the typical cheerful happy-go-lucky girl and like I said, it was manageable at best. I get into hobbies if I ever find myself falling into it. I think he doesn't understand that depression does creep up on you. You think it's gone and then it's back again. I try to keep my mind healthy. I allow myself to be

Higher entropy

- 131 Well to be fair moving across continents can be very complicated. Logistically and emotionally. Has he specified what the reasons are? Sure, if it's that he's not sure enough about you for eg, then you should keep your space. But if it's financial, visas, job, family ties, the specific country etc then you need to talk about the specifics and decide if there's a compromise, when that would work and if you're both willing to wait that long.
- 132 Seems a bit flowery. While Ellison may be the inspiration of the signs, the men didn't assert their presence or make themselves visible by deleting "invisible". The phrase stands on its own & I don't think you could have come up with a sign that more succinctly stated their motivation for striking. "I am a Man" says it all. Very powerful image.
- 133 The lyrics to Starman by David Bowie seem incredibly psychedelic to me. Has anyone heard of Bowie mentioning this direct link? His use of psychedelics

Even higher entropy

```
170 Setting aside the "is Mathematics Universal?" debate, there is one number
    system that could be considered universal, Unary. As for radio signals, there
    is again a way of transmitting that one could consider Universal: On/Off.-\_-\_-\_
    \_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_-\_
171     wait = WebDriverWait(driver, 30)     elems = wait.until(
        EC.presence_of_element_located((By.CSS_SELECTOR,
        'div.account-summary-tile:nth-child(3) > div:nth-child(1) >
        div:nth-child(1) > div:nth-child(2) > div:nth-child(1) > div:nth-child(1) >
        a:nth-child(1)')))
```

Very highest entropy

```
263 #####&#009;#####&#009;####&#009; [**Makyo**](https://
en.wikipedia.org/wiki/Makyo): [](#sfw) ---&gt;&gt;The term __makyo__ (魔境,
*makyō* ?)
264 I wasn't on a bike.
265 *---^Interesting: [^Stokes ^stream ^function](https://en.wikipedia.org/wiki/
Stokes_stream_function) ^| [^Vector ^potential](https://en.wikipedia.org/wiki/
Vector_potential) ^| [^Navier–Stokes ^equations](https://en.wikipedia.org/wiki/
Navier%E2%80%93Stokes_equations) ^| [^Milne–Thomson ^circle ^theorem](https://
en.wikipedia.org/wiki/Milne–Thomson_circle_theorem) ^Parent ^commenter ^can
[^toggle ^NSFW](/message/compose?
266 Multiple sources say otherwisehttp://i.imgur.com/2QKpATa.png
267 (https://www.google.pl/search?tbm=isch&tbs=simg:CAQSxQEawgELEKjU2AQaAAwLELCM
pwganAEK0ggCEhSqFLghtyGSFJAU5xrRCpMUoRSgFBogNtf-7Sv5HEgz2SvLPqtgK59uBDP92ZFFNpVq
Gj0biKcKXggDEijiFZ0e_1x3qF0sUwRbgFZwe7RThFZA_1nDiaP_1A3mzjxN9YpljiBK_1YpGjBeyoIg
o_1d-G38rD7Ehm1yJvSU369Zi09s5yfmbuWmyWmm5zfvuIpd10oirNQuRdzcMCxC0rv4IGgoKCAgBEgS
qrr4kDA&sa=X&ei=3jzAVI2vEau1ygPD0IGwCg&ved=0CB4Q2A4oAQ&gws_rd=cr
,ssl#imgdii=)
```

Shannon's definition of **entropy**

Theorem 2: The only H satisfying the three above assumptions is of the form:

$$H = -K \sum_{i=1}^n p_i \log p_i$$

where K is a positive constant.

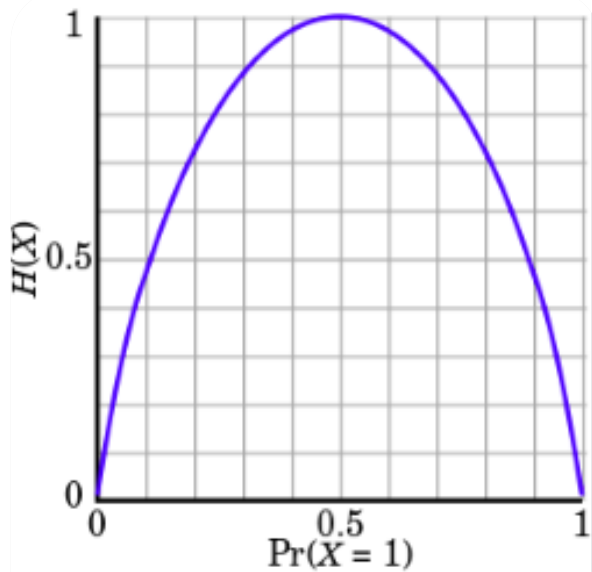
This theorem, and the assumptions required for its proof, are in no way necessary for the present theory. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions is given in the next chapter. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions is given in the next chapter.

- .. minimum number of bits needed to encode a string of symbols, based on the frequency of the symbols ..
- “The form of H will be recognized as that of **entropy** as defined in certain formulations of statistical mechanics where p_i is the probability of a system being in cell i of its phase space.”

- Bell System Technical Journal, 1948

“A Mathematical Theory of Communication”
Shannon, C.E.

Shannon's definition of entropy



Entropy $H(X)$ (i.e. the expected surprisal) of a coin flip, measured in bits

.. minimum number of bits needed to encode a string of symbols, based on the frequency of the symbols ..

“The form of H will be recognized as that of entropy as defined in certain formulations of statistical mechanics where p_i is the probability of a system being in cell i of its phase space.”

Bell System Technical Journal, 1948

“A Mathematical Theory of Communication” Shannon, C.E.

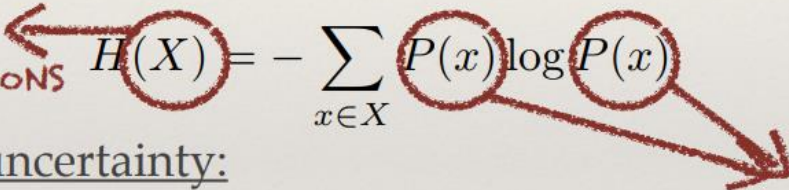
Quantifying a word's information content

- ❖ **entropy**: how uncertain we are about what is being communicated at a certain point
(it decreases with each upcoming word)

POSSIBLE
INTERPRETATIONS

$$H(X) = - \sum_{x \in X} P(x) \log P(x)$$

$P(x|w^{t_1})$
after the first w



- ❖ maximal uncertainty:
all interpretations are equally probable
- ❖ minimal uncertainty:
we are certain about one particular interpretation ($P(x) = 1$)

Shannon Entropy, very bad entropy and other measures

- Just counting how common a sentence is
- Shannon entropy
- Compression entropy (online compression)
- “Very bad entropy” (PySmaz)

Semantic ‘Entropy – RNN Entropy

2.1. Model architecture

Fig. 1 presents the architecture of the recurrent neural network (RNN) that was used as the probabilistic language model for estimating word-surprisal and entropy-reduction values. This network is not proposed as a cognitive model; rather, it serves as a tool for obtaining the required word-information measures with several advantages over alternative models. For one, RNNs process more efficiently than phrase-structure grammars (or other structure-assigning models), which is of particular importance for computing entropy. Also, they can be trained on unannotated sentences (i.e., word strings instead of tree structures). In addition, RNNs have been shown to estimate surprisal values that fit reading times better than do grammar-based surprisal estimates (Frank & Bod, 2011). This was also demonstrated by Fernandez Monsalve et al. (2012), using the very same

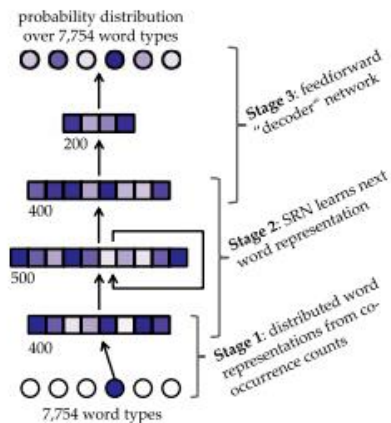


Fig. 1. Architecture of neural network language model, and its three learning stages. Numbers indicate the number of units in each network layer. Reproduced from Fernandez Monsalve et al. (2012).

Semantic 'Entropy – W2V Entropy'

```
# the log likelihood of this 'sentence'
# under the given w2v representation
# calculate the 'entropy' as the average semantic
# difference between words – larger distances = more 'entropy'
words = utils.simple_preprocess(txt)
# words = [word for word in sentence.lower().split()]
out_of_vocab = 0
in_vocab_words = []
for word in words:
    if model.vocab.has_key(word):
        in_vocab_words.append(word)
    else:
        out_of_vocab += 1

distances = {}
for aWord in in_vocab_words:
    for bWord in in_vocab_words:
        if not (aWord == bWord):
            key = two_word_key(aWord, bWord)
            # we only need to calculate an upper right triangle
            if not (key in distances):
                distances[key] = 1 - abs(model.similarity(aWord, bWord))
            # sysout("distance " + key + " = " + str(distances[key]))

# get average
total_distance = float(sum([val for val in distances.values()]))
```

Sentence online with highest W2V Entropy?

James Joyce – Ulysses - for the win

Its universality: its democratic equality and constancy to its nature in seeking its own level: its vastness in the ocean of Mercator's projection: its unplumbed profundity in the Sundam trench of the Pacific exceeding 8000 fathoms: the restlessness of its waves and surface particles visiting in turn all points of its seaboard: the independence of its units: the variability of states of sea: its hydrostatic quiescence in calm: its hydrokinetic turgidity in neap and spring tides: its subsidence after devastation: its sterility in the circumpolar icecaps, arctic and antarctic: its climatic and commercial significance: its preponderance of 3 to 1 over the dry land of the globe: its indisputable hegemony extending in square leagues over all the region below the subequatorial tropic of Capricorn: the multiseular stability of its primeval basin: its luteofulvous bed: its capacity to dissolve and hold in solution all soluble substances including millions of tons of the most precious metals: its slow erosions of peninsulas and islands, its persistent formation of homothetic islands, peninsulas and downwardtending promontories: its alluvial deposits: its weight and volume and density: its imperturbability in lagoons and highland tarns: its gradation of colours in the torrid and temperate and frigid zones: its vehicular ramifications in continental

The background of the slide is a photograph of the Guggenheim Museum in Bilbao, Spain, showing its iconic curved, metallic facade. A yellow puzzle piece icon is centered below the text.

Other experiments

Content Shaping

```
# foreach word in sentence
#   if word < threshold for any words in pref, score + a little
#   if word < threshold for any words in anti, score - a little
def score(self, sentence):
    score = 0
    pref_factor = float(1) / len(self.pref_words)
    anti_factor = float(1) / len(self.anti_words)
    words = simple_preprocess(sentence)
    counted = defaultdict(int)
    for word in words:
        counted[word] += 1
        if counted[word] > 50:
            # to protect against those crazy word repeating lunatics
            return -10
    for pref in self.pref_words:
        if self.score_above_threshold(word, pref):
            if counted[word] < 6:
                score += pref_factor
                break
    for anti in self.anti_words:
        if self.score_above_threshold(word, anti):
            if counted[anti] < 6:
                score -= anti_factor
                break
    return score
```

Semantic Resonance

136 I couldn't imaging how that feels.
137 I wouldn't simplify it that much.
138 i wouldn't define it that way.
139 I wasn't debating you or anything.
140 Doesn't raccoon have it to?
141 Amused mastery all the way.
142 I didn't contour there (I never do).
143
144 My daughters friend had leukemia.
145 My mother thinks I have aspergers.
146 my friends dad is a geologist.
147 My grandparents left before WWI.
148 My pitbull's name is "Piggy."
149 Your fiancé's idea is irresponsible.
150 One friend said I was a succubus.
151
152 it is inconsequential to me.
153 That is unconscionable to me.
154 This seems disingenuous to me.
155 It seems abysmal to me.
156 It's not cyclical for me.
157 It seems improper to me.
158 It's mind boggling to me :(

Rhyming

```
145856  
145857 Wow, this is amazing!  
145858 Shit, that's amazing!  
145859 > ...that was closing  
145860  
145861 Wow, very unsettling!  
145862 Low-quality trolling.  
145863 How is that trolling?  
145864  
145865 Wow, you're a wizard.  
145866 Gotta ask Blizzard  
145867 I am a lizard???  
145868
```



On working with artists





Working with **artists**

- Don't Dunning Kruger it
 - realize how much you don't know
- Don't show algorithms without being cognizant of other aesthetics
 - wait a bit more till ready?
- “Open source” ?
 - can make some artists twitch
- Respect
 - is spelled R E S P E C T

.. we live in a world in which the question of agency is highly disputed, in which our access to choice is circumscribed by the framing of political discourse, by the **simultaneous cornucopia and contraction** of “options” in our consumer paradise, and, indeed, by algorithms, which filter, consolidate and display certain possibilities while rendering all others invisible.

- Annie Dorsen, 2012



“



Thanks!

Any **questions** ?

You can find me at

- @utungu
- miles@goodbright.nz
- www.goodbright.nz